Computer Vision: Classification of Skin Lesions from the HAM10000 Dataset

Henri Antikainen
Bailey Forster
Tristan Freiberg

Project Overview and Objectives

Melanoma can affect anyone and early detection is a crucial factor affecting survival rates.
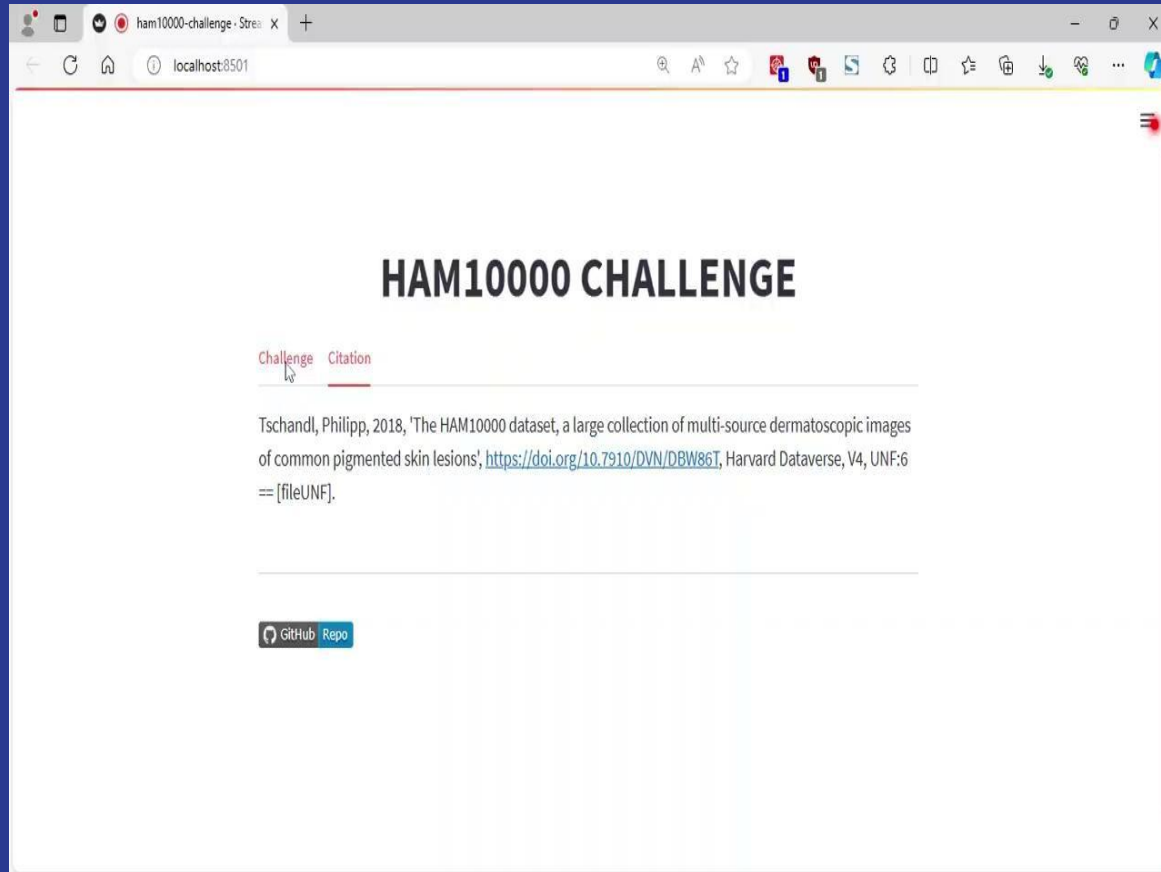
Machine learning models could assist trained healthcare professionals in screening for skin cancer.

Goals:

- Classify images of skin lesions using convolutional neural networks (ResNet and EfficientNet architectures) on the HAM10000 dataset.
- Test average human ability to identify melanoma against a CNN model

Streamlit App Challenge

Tests users' ability to classify skin lesions as moles or melanoma.

Compete against one of our fine-tuned machine learning models

The Human-Against-Machine 10000 Dataset

- Contains ~10,000 images of 7,470 distinct skin lesions

- 7 classes
  - Cancerous: e.g. melanoma, basal cell carcinoma
  - Benign: e.g. nevi (moles)

- Challenge: Highly imbalanced dataset
  - Melanoma: <10%, Nevi: >70%
  - Cancerous classes have much lower frequency

- Performance metric: Balanced accuracy = (sensitivity+specificity)/2

Preprocessing

- 5 classes: 3 cancerous types, nevi (moles), and other benign lesions
- Train-test split: used stratified split to assign each lesion ID to the training or test set while keeping proportions consistent
- Handling multiple images of a lesion:
    1. Keep all images
    2. Randomly select one image per lesion
- Balancing the dataset: undersampled the nevi and oversampled smaller classes, so that each class was represented by 2,000 lesions.

Model Training and Parameters

Trained two CNN models: Resnet18 and EfficientNet
- Cross-entropy loss functions and Adam optimization
- Default 10 epochs

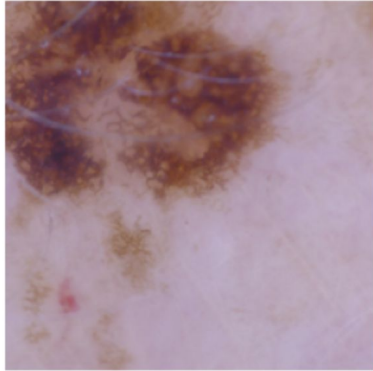We experimented with several parameters and methods of augmentation:
- Unbalanced vs. balanced dataset
- Unfreezing layers: all layers or last only
- Transformations: random crop, rotation, contrast, brightness, hue

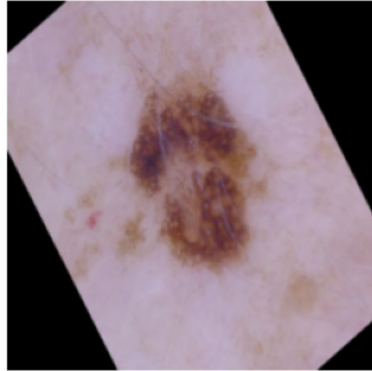Post-training: Selecting classification thresholds
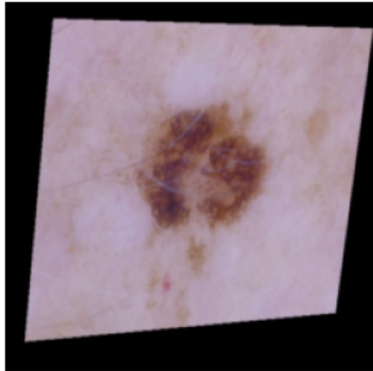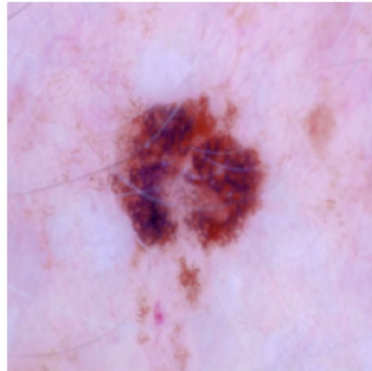
Augmentation
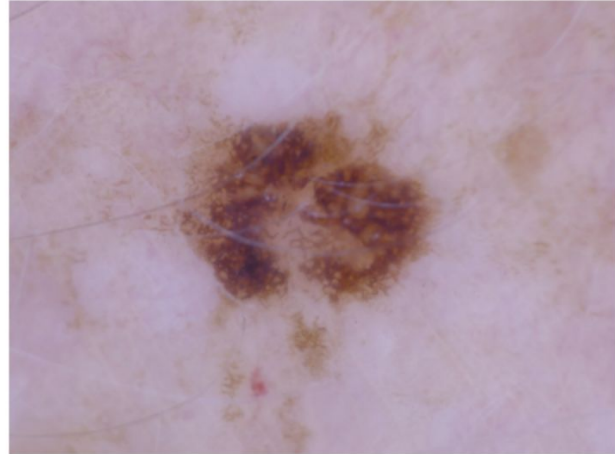
Results: Resnet18

Model	Class distribution	Images per lesion	Layers unfrozen	Transformations	BACC	Melanoma Recall
02	unbalanced	1	all	-	0.381	0.169
03	unbalanced	1	last	-	0.481	0.247
04	balanced	1	last	-	0.554	0.403
05	balanced	all	last	random crop	0.695	0.740
06	balanced	all	last	random crop + color jitter	0.595	0.675
07	balanced	all	last	random rotation	0.663	0.578
Modified prediction threshold with bias towards melanoma:						
05	balanced	all	last	random crop	0.700	0.773

Results: EfficientNet

EfficientNet: often outperforms ResNet in benchmarks, systematic approach to scaling

Model	Class distribution	Images per lesion	Layers unfrozen	Transformations	BACC
RN05	balanced	all	last	random crop	0.70
ENB0	balanced	all	last 2	random crop	0.73

Future Work

Improving our models:

- Incorporate dropout regularization to mitigate overfitting
- Create custom transformations that address artifacts in images, such as markings from dermatologists, for improved classification accuracy
- Include metadata (e.g. age, sex, localization of lesion) as model features

Other challenges/ideas:

- Representation of different ethnicities is a concern in much of the available skin cancer data
- Use these results as pretrained model and finetune with a more diverse dataset

Sources & Acknowledgements

Thank you to Lindsay Warrenburg and Erdös Institute for organizing the deep learning bootcamp

Sources:

HAM10000 dataset: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T

[1] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, Allan Halpern: "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)", 2018; https://arxiv.org/abs/1902.03368

[2] Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 180161 doi:10.1038/sdata.2018.161 (2018).